Article ID: 1002-1175(2010)01-0017-10

A three-stepwise robust statistical method for outlying rainfall observation*

ZHAO Chao[1,2†], HONG Hua-Sheng[2], ZHU Mu-Lan[1]

(1 Water Resources and Environmental Institute, Xiamen University of Technology, Xiamen 361005, China;
2 State Key Laboratory of Marine Environmental Science, Environmental Science Research Center,
Xiamen University, Xiamen 361005, China)

(Received 6 November 2008; Revised 3 September 2009)

Zhao C, Hong H S, Zhu M L. A three-stepwise robust statistical method for outlying rainfall observation[J]. Journal of the Graduate School of the Chinese Academy of Sciences, 2010, 27(1): 17-26.

Abstract A three-stepwise robust statistical method combining the robust statistical theory with distribution features of rainfall for detection of outliers in telemetry system is described. The proposed robust statistical method adopts the Tukey fence insensitive to outliers as identification bounds and presents a three-stepwise pattern to adapt the distribution of rainfall data. Moreover, the modified method based on dividing precipitation data into several groups further improves detection efficiency. The results show that the new method is suitable to the hydrological need.

Key words telemetry system; outlier; Tukey fence; distribution feature; three-stepwise robust detection

CLC P338

1 Introduction

Rainfall is the primary input to most hydrologic models, therefore precision of rainfall data can have a key effect on accuracy of hydrologic forecasting and simulation.

Historically manual observations at rainfall stations have been the main source of rainfall data. More recently, automatic acquisition on telemetric network has played an increasingly important role in technologically developed countries, particularly with regard to provision of data in real time. It is obvious that the rainfall data obtained by telemetry system is characterized as accuracy and frequency. However, it is noted that in telemetric rainfall data there often are outliers resulting from instrument malfunctioning and false signal acquisition because of signal leak, collision and disturbance during signal transmission, in addition to unavoidable random errors normally distributed with zero mean and a small variance. The outliers have an unknown distribution with a much bigger variance, and appear to be inconsistent with the remainder of the data set[1-2] and relative large in magnitude.

* 国家自然科学基金(50909084)、福建省自然科学基金(2009J05107)和校级引进人才项目(YKJ08015R)资助

†E-mail: zhaochao@xmut.edu.cn

In order to limit the influence of outliers on forecasting and simulation, it is essential to check and find out these outlying observations before they enter into the hydrologic models.

The detection of outliers has been tackled by a variety of methods in hydrology. These can be classified into two kinds. One is called statistical test based on some hypotheses. For example, Bulletin 17B prevalently used to analyze annual maximum flood records for the purpose of determining flood flow-frequency curves is based on the principle of hypothesis testing with the underlying assumption of log Pearson Type III probability distribution[3-4]. The method is to establish a limit at the mean $\pm k$* SD(standard deviation) where k is the critical deviate at the selected level of significance. The values outside the limit are considered to be outliers and rejected. But in many recent literature[5-9], the conclusions that even one outlier seriously influences the mean, SD and limits set, and the limits are deviated toward the outlier so that the outlier which could be masked have been proposed. The other is called comparison test based on functional interpolation and/or stochastic estimation. The approach used by the US Environmental Protection Agency (US-EPA) employing Mahalonobis distance for the detection of outliers belongs to this class. Outliers typically have larger Mahalonobis distances than "good" observations. The similar results as above are gained that outliers evidently affect the covariance matrix and Mahalonobis distances so that outliers are masked. All the methods mentioned above are based on the sample mean, standard deviation and covariance matrix, which suffer from the heavy influence of outliers. Therefore, a new detecting method which is more tolerant to outliers is needed.

Otherwise rainfall varies temporally and spatially in a watershed system. In order to correctly detect the outlying observations, the detection method has to depend on the statistical feature of the variety.

Considering synthetically those constrains, this paper makes an attempt to combine the robust statistical approach and distribution characteristics to increase the reliability for the detection of outliers. A three-stepwise robust statistical detection method of outliers is proposed. The robust approach[7-10] based on inter-quartile range (IQR) is more tolerant of outliers. The three-stepwise pattern agrees with the different distribution statistical features of hourly rainfall (HR), hourly rainfall deviation from hourly areal mean rainfall (HRD), and hourly simulation error of hourly rainfall simulation (HSE).

This paper is organized as follows. Firstly, a brief review of statistical property of rainfall distribution is given in section 2. Then the proposed method in detail is presented in section 3. A case study is conducted to demonstrate the proposed method in section 4, and section 5 contains some conclusions.

2 Statistical features

For a watershed system, rainfall varies temporally and spatially. How to determine the outliers depends on the features of the distribution.

The data used in this study is taken from hourly manual rainstorm data of 43 raingauges from 1988 to 1997 in Qilijie basin of Fujian province of China to analyze the statistical features of distribution. The distributions of HR, HRD and HSE are displayed in Fig. 1, Fig. 2, and Fig. 3 respectively, where only distribution for Wuyigong station is depicted, since identical results are observed for other stations.

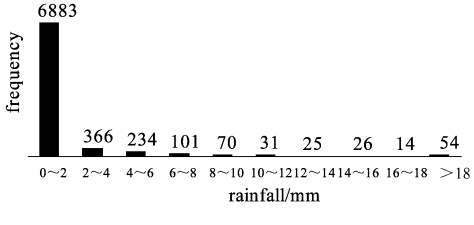

Fig. 1 HR distribution of Wuyigong station

HR distribution is very complex, and depends upon air current, topography, elevation, geographic position, and some other factors. HR always ranges from 0 to maximum precipitation(P_{max}). The distribution graph has a single peak, and 92.9 percent of HR locate in [0, 4mm] which is the high frequency zone. The mean square deviation displaying dispersion degree of HR is 2.88.

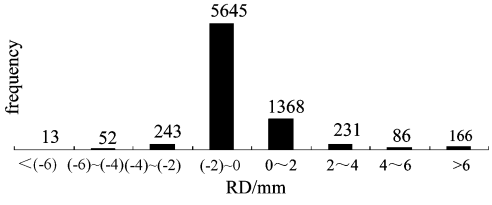

Fig. 2　RD frequency distribution of Wuyigong station

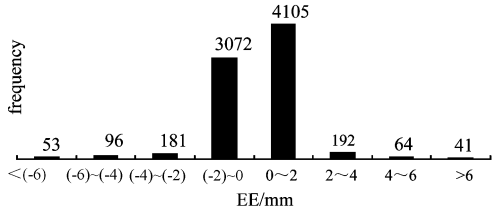

Fig. 3　EE frequency distribution of Wuyigong station

HRD is got by following equation

$$rd_{ij} = P_{ij} - P_{\mathrm{mean},j}, \tag{1}$$

where P_{ij} is hourly rainfall of the ith raingauge and the jth time-interval, P_{mean} is hourly areal mean rainfall computed by arithmetic mean rainfall of all raingauges except the ith one.

The 95.9 percents of HRD locates in [(-4) ~ 4mm] where is the high frequency zone. The mean square deviation of HRD is 2.24.

Rainfall simulation is a difficult task. This paper selected a simple model of quadric by means of longitudes and latitudes of raingauges and P_{mean}, to simulate HR. HSE is achieved

$$ee_{ij} = P_{ij} - P_{\mathrm{sim}ij}, \tag{2}$$

where P_{sim} is the simulated HR of the ith raingauge and the jth time-interval.

HSE locate in [(-4) ~ 4mm]. The mean square deviation of HSE is 2.18.

For all raingauges the distributions of HR, HRD and HSE have the following features: Firstly, the distributions are limited in [0, maximum], and the distribution ranges are less and less from HR to absolute HRD to absolute HSE. For example, for Wuyigong station, the distribution ranges of precipitation, absolute HRD and absolute HSE are [0, 36.9mm], [0, 30.7mm], and [0, 28.6mm], respectively. For 43 raingauges, the ranges of three distributions are [0, 65.2mm], [0, 58.1mm], and [0, 50mm], respectively. Secondly, for all raingauges, the mean square deviations of the three distributions are 2.38, 2.10, and 1.94 respectively, which shows the three distributions are closer and closer.

3　Methods

3.1　Traditional methods

Traditional methods are based on the estimates of the mean, standard deviation and covariance matrix which can be seriously distorted by even one outlier. Thus, it is unreasonable to use mean, standard deviation and covariance matrix, which are relavant to the outliers for establishment of the targets to detect outliers.

3.2　Robust statistical method

If $X_n = \{x_1, x_2, \cdots, x_n\}$ is a univariate data set. Let $X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$ denote the ascending sequence. The p-quantile of data set has the form

$$X_{(p)} = \begin{cases} X_{(np)}, & np \in \text{int}, \\ X_{[(np)+1]}, & np \notin \text{int}, \end{cases} \tag{3}$$

where $[\cdot]$ denotes the integer part of the number.

The 1/4 quantile ($X_{(1/4)}$) is called lower quartile which has 1/4 data$< X_{(1/4)}$, and 3/4 quantile ($X_{(3/4)}$) is called upper quartile which have 1/4 data$> X_{(3/4)}$.

Obviously, the outliers are unusually large or small, and arranged at both end of the ordered sequence. So the quantiles, for example $X_{(1/4)}$ and $X_{(3/4)}$, are less related to outliers than $\bar{X}$ and σ. Therefore, we

selected the robust statistical method based on quantiles as the detection tool.

The difference between $X_{(1/4)}$ and $X_{(3/4)}$ is named inter-quartile range (IQR). The Tukey fence[8-11] values are in the range from $X_{(1/4)} - \beta * IQR$ to $X_{(3/4)} + \beta * IQR$. The data outside the Tukey fence are considered to be outliers. $X_{(1/4)}$ and $X_{(3/4)}$ are less related to outliers, so the Tukey fence is not skewed by the outliers. β is generally estimated by trial-and-error method considering level of significant. If the data come from the normal distribution and level of significance is 1%, $\beta = 1.5$ that the fence contains 99.5% of the data. We select 10% level of significance for hydrologic rainfall data where a high probability of mixed distributions exists. This is the robust statistical method of detecting outliers.

In order to ensure the detection efficiency of robust statistical method, we incorporate rainfall distribution features into the robust statistical method, and propose a three-stepwise robust statistical method to detect the outliers in telemetric rainfall sets.

The three-stepwise robust statistical method is showed as follows.

Above all, the upper quartiles, lower quartiles, IQRs and the Turkey fences of the HR, HRD and HSE are computed.

Step 1 Judge whether the real-time hourly rainfall is in Tukey fence of HR or not. If rainfall observations are not in, they are questionable, and then end detecting. Otherwise, enter into next step.

Step 2 Judge whether the real-time HRD is in its Tukey fence or not. If the data are not in, they are questionable, and then end detecting. Otherwise, enter into next step.

Step 3 Judge whether the real-time HSE is in its Tukey fence or not. If the data are not in, they are questionable, and then end detecting.

Detection is only part of an analysis of outliers. The statistical detection should be used to alert the analyst to investigate the cause of outlying rainfall observations. If reasons are found for the outliers, they should be treated accordingly. If no cause can be determined, Grubbs[12] proposes the outliers could be retained. In this paper, the main goal is to attain the detecting purpose of the three-stepwise method, the analysis of the cause will be done in future. The outlying observation is assumed by the outliers in this paper.

Otherwise in the process of computing the Turkey fences, we find that rainfall data of different magnitudes have large differences in $X_{(1/4)}$, $X_{(3/4)}$ and IQRs. So a modified detection method is proposed to further improve the detection efficiency.

3.3 Modified method

To consider the different magnitudes of rainfall sets, we divide the areal mean precipitation into several groups. For the every group, the upper quartiles, lower quartiles and IQRs of the three distributions are recalculated. The modified method is showed as follows:

Step 1 Compute the real-time hourly areal mean precipitation (P_{mean}), and select the corresponding group and Tukey fence of HR. Judge whether the real-time hourly rainfall are in Tukey fence or not. If the data are not in, they are outliers, and then end detecting. Otherwise, enter into next step.

Step 2 Recalculate the P_{mean}, and select the corresponding group and Tukey fence of HRD. Judge whether the data are in Tukey fence or not. If the data are not in, they are outliers, and then end detecting. Otherwise, enter into next step.

Step 3 Recalculate the P_{mean}, and select the corresponding group and Tukey fence of HSE. Judge whether the data are in Tukey fence or not. If the data are not in, they are outliers, and then end detecting.

All the outlying rainfall observation is replaced by the average rainfall value of other stations.

4 Experimental results

HR data used in this section are manual rainfall observation in 1998 totaled 23000 pieces. They only contain random errors which could enter into hydrologic models.

The performance of the proposed robust statistical method is demonstrated using synthetically generated data sets. The reason for using synthetic data is that the error structure was known and could be varied to test different hypotheses.

We add the following error distribution to the rainfall data set of single raingauge in 1998, in order to simulate the outlying distribution of telemetric system:

$$e_i = \begin{cases} re_{\max}, & i = \text{int}(i/L)L; \\ 0, & i \neq \text{int}(i/L)L, \end{cases} \tag{4}$$

where r is a random number; $e_{\max}$ is a constant that controls the maximum of e. L is the frequency of outliers. Adjusting $e_{\max}$, outliers of different sizes can be generated.

4.1 Performance of robust statistical method

We add the outliers using equation (4) to precipitation data of one, two and five raingauges.

The Tukey fences of HR, HRD and HSE distribution is calculated. In order to contain 90% of the values, β is 2.5 by means of trial trial-and-error method.

In evaluation of the robust statistical detection method, the following two performance measures[13] are used in this study. One is detection efficiency (ee), the other is detection mistake (m) to detect true data as outliers.

$$ee = \frac{n_d}{n_o}, \tag{5}$$

where n_d is the number of outliers correctly detected, n_o is the number of outliers.

$$m = \text{the number of outliers wrongly detected} \tag{6}$$

The detection efficiency of the proposed method is showed in Table 1. As a whole, the proposed method is proper to detect the outliers for telemetric rainfall sets, and the mean detect efficiency is above 0.6. For single raingauge, the detection efficiency enhanced as $e_{\max}$ increased. $ee1$ is 0.02 for 10mm of $e_{\max}$ and 0.96 for 500mm. Among the three steps, the detection efficiency of the first step is the most obvious, and that of the third is least. For example, for one raingauge and $e_{\max}$ = 70mm, the detection efficiency of the first, second and third step is 0.5, 0.18, 0, respectively. For two and five raingauges, similar results could be obtained. As the number of stations contaminated is on the increase, the efficiency of first step reduces, which of the second and third step advances. It means the performances of the three steps compensate one another. However, the changes of the detection efficiency are not large.

The detection mistake is given in Table 2. It is obvious that the highest detection mistake occurred in the second step, which indicates that the Tukey fence of the second step is not relevant to true data. The changes of detection mistake are little as $e_{\max}$ increased. As the number of raingauges contaminated is on the increase, the detection mistake of the first step reduces, and that of the second and third step increases in tendency.

The performance of the three-stepwise robust statistical detection method is measured by the r_ij (i= 1, 2, 3; j= 1, 2, 5)[14] computed by follows:

$$r_1j = 1 - \frac{\sum |\hat{e}_1j|}{\sum |\hat{e}_0j|}, \tag{7}$$

$$r_2 j = 1 - \frac{\sum |\hat{e}_2 j|}{\sum |\hat{e}_1 j|}, \tag{8}$$

$$r_3 j = 1 - \frac{\sum |\hat{e}_3 j|}{\sum |\hat{e}_2 j|}, \tag{9}$$

$$r_t j = 1 - \frac{\sum |\hat{e}_3 j|}{\sum |\hat{e}_0 j|}, \tag{10}$$

Table 1 Detection efficiency of robust statistical method

e_{max} /mm	$e1$	$e_1 1$	$e_2 1$	$e_3 1$	$e2$	$e_1 2$	$e_2 2$	$e_3 2$	$e5$	$e_1 5$	$e_2 5$	$e_3 5$
10	0.02	0	0.02	0	0	0	0	0	0	0	0	0
20	0.06	0	0.06	0	0.13	0	0.13	0	0.04	0	0.04	0
30	0.36	0	0.36	0	0.43	0.01	0.41	0.01	0.32	0.02	0.30	0
40	0.68	0.32	0.36	0	0.53	0.14	0.39	0	0.53	0.17	0.34	0.02
50	0.66	0.32	0.34	0	0.69	0.40	0.29	0	0.60	0.30	0.29	0.01
60	0.76	0.54	0.22	0	0.70	0.48	0.21	0.01	0.71	0.50	0.21	0
70	0.68	0.50	0.18	0	0.63	0.46	0.17	0	0.76	0.54	0.22	0
80	0.72	0.58	0.14	0	0.81	0.67	0.14	0	0.80	0.64	0.16	0
90	0.78	0.60	0.18	0	0.85	0.64	0.21	0	0.79	0.65	0.14	0
100	0.88	0.82	0.06	0	0.80	0.59	0.21	0	0.81	0.65	0.15	0.01
200	0.92	0.82	0.10	0	0.95	0.87	0.08	0	0.91	0.87	0.04	0
500	0.96	0.96	0	0	0.98	0.95	0.03	0	0.98	0.95	0.03	0
mean	0.62	0.46	0.17	0	0.63	0.43	0.19	0.00	0.60	0.44	0.16	0.00

注: $e1$, $e2$, $e5$: detection efficiency of data of 1, 2, 5 stations containing outliers using equation (4), respectively; Subscript 1, 2, 3: the first, second, third step, respectively.

Table 2 Detection mistake of robust statistical method

e_{max} /mm	$m1$	$m_1 1$	$m_2 1$	$m_3 1$	$m2$	$m_1 2$	$m_2 2$	$m_3 2$	$m5$	$m_1 5$	$m_2 5$	$m_3 5$
10	12	1	11	0	12	1	11	0	13	1	12	0
20	12	2	10	0	12	1	11	0	12	2	10	0
30	12	2	10	0	12	2	10	0	13	1	9	3
40	12	1	11	0	12	2	10	0	11	1	8	2
50	11	1	10	0	11	1	10	0	11	1	8	2
60	11	1	10	0	12	1	11	0	11	1	10	0
70	11	1	10	0	11	1	10	0	11	1	10	0
80	11	1	10	0	12	2	10	0	11	1	10	0
90	11	1	10	0	11	1	10	0	11	1	10	0
100	11	1	10	0	11	1	10	0	12	1	11	0
200	11	1	10	0	11	1	10	0	11	1	10	0
500	11	1	10	0	11	1	10	0	13	1	12	0

注: $m1$, $m2$, $m5$: detection mistake of data of 1, 2, 5 stations containing outliers using equation (4), respectively.

where e_0 is the outlying error generated by equation (4); e_1, e_2, e_3 are the errors of after the first, second and third step detection, respectively; r_1, r_2, r_3 are the performance of the first, second and third step, respectively; r_t is the total performance of the method.

It is noted that the performances are resulted from the detection efficiency and mistake. The performances of increasing e_{max} from 10mm to 500mm are displayed in Table 3.

Table 3　Performances of robust statistical detection method

e_{max}/mm	$r_1$1	$r_2$1	$r_3$1	r_t1	$r_1$2	$r_2$2	$r_3$2	r_t2	$r_1$5	$r_2$5	$r_3$5	r_t5
10	-0.151	-0.859	0	-1.141	-0.074	-0.451	0	-0.558	-0.027	-0.174	0	-0.206
20	-0.087	-0.333	0	-0.448	-0.031	0.015	0	-0.016	-0.015	-0.020	0.007	-0.029
30	-0.056	0.246	0	0.204	-0.022	0.447	0.019	0.445	0.004	0.431	-0.030	0.417
40	0.421	0.353	0	0.625	0.221	0.498	0	0.609	0.274	0.522	0.027	0.662
50	0.439	0.316	0	0.616	0.581	0.436	0	0.763	0.488	0.512	-0.005	0.749
60	0.703	0.099	0	0.733	0.693	0.336	0.028	0.802	0.716	0.490	0	0.855
70	0.711	0.052	0	0.726	0.713	0.207	0	0.772	0.751	0.521	0	0.880
80	0.783	-0.113	0	0.758	0.845	0.190	0	0.874	0.842	0.431	0	0.910
90	0.816	0.001	0	0.816	0.821	0.372	0	0.887	0.848	0.384	0	0.907
100	0.917	-0.583	0	0.869	0.818	0.389	0	0.889	0.869	0.431	0.040	0.929
200	0.950	-0.328	0	0.934	0.962	-0.069	0	0.959	0.971	0.031	0	0.972
500	0.991	-1.728	0	0.974	0.990	-0.645	0	0.984	0.991	-0.078	0	0.99
mean	0.536	-0.240	0	0.472	0.543	0.144	0.004	0.618	0.559	0.290	0.003	0.670

When data of only one raingauge contained outliers, the total performances increases as the e_{max} increases: r_t1 is -1.151 for 10mm and 0.991 for 500mm. For two and five raingauges, similar results are obtained. It means the performances of the robust statistical method are obvious with an increase in e_{max}. Among the three steps, the performances of the first step are greatest, and the performances of the third step are least. For example, for one station of e_{max} = 70mm, the performances of the first, second and third step are 0.711, 0.052, 0, respectively. Moreover, as the number of raingauges contaminated increases, the performances of the second, third step are more and more obvious. These results agree with those of detection efficiency and mistake.

The negative value in Table 3 results from the detection mistake. When e_{max} is small, the detection efficiency is also small, causing the response of detection mistake is obvious. When e_{max} increases, the detection mistake still occurs, however its impact is concealed by the response of the detection efficiency.

4.2　Performances of the modified method

The hourly rainstorm data containing outlying errors in 1998 are divided into five grades, i.e., [0~1mm], (1mm~2mm], (2mm~4mm], (4mm~6mm], (>6mm). The Tukey fences of every group are calculated.

The detection efficiency of the modified method is shown in Table 4. The mean detection efficiency is above 0.68. The detection efficiency increases as e_{max} increased, and it reduces as the number of raingauges contaminated increases. The most obvious efficiency occurs in the first step. Moreover, as the number of raingauges contaminated increases, the efficiency of the second and third step is larger and larger. Comparing the results in Table 1 with Table 4, the detection efficiency of the modified method is larger, particularly when e_{max} is small.

Table 4 Detection efficiency of the modified method

e_{max}/mm	$E1$	$E_1 1$	$E_2 1$	$E_3 1$	$E2$	$E_1 2$	$E_2 2$	$E_3 2$	$E5$	$E_1 5$	$E_2 5$	$E_3 5$
10	0.100	0	0.100	0	0.060	0	0.060	0	0.070	0.004	0.068	0
20	0.460	0.360	0.100	0	0.520	0.340	0.180	0	0.410	0.192	0.184	0.036
30	0.580	0.560	0.020	0	0.650	0.500	0.150	0	0.460	0.260	0.160	0.04
40	0.740	0.720	0.020	0	0.670	0.570	0.100	0	0.660	0.360	0.264	0.036
50	0.760	0.740	0.020	0	0.730	0.650	0.080	0	0.680	0.396	0.228	0.052
60	0.860	0.860	0	0	0.780	0.680	0.100	0	0.720	0.464	0.228	0.028
70	0.780	0.780	0	0	0.780	0.660	0.120	0	0.780	0.488	0.268	0.024
80	0.840	0.800	0.040	0	0.840	0.750	0.090	0	0.830	0.516	0.276	0.036
90	0.820	0.800	0.020	0	0.890	0.800	0.080	0.010	0.830	0.512	0.284	0.036
100	0.900	0.900	0	0	0.830	0.760	0.070	0	0.820	0.504	0.272	0.048
200	0.880	0.880	0	0	0.950	0.850	0.100	0	0.930	0.708	0.196	0.028
500	1.000	1.000	0	0	0.990	0.920	0.070	0	0.970	0.904	0.068	0
mean	0.727	0.700	0.027	0	0.724	0.623	0.100	0.001	0.680	0.442	0.208	0.030

注: $E1, E2, E5$: detection efficiency of data of 1, 2, 5 stations containing outliers using the modified method and equation(4), respectively.

The detection mistake of the modified method is obtained in Table 5. As the number of raingauges contaminated increases, the detection mistake of the first step reduces, and which of the second and third step increases. The result agrees with detection efficiency. Comparing the results in Table 2 and Table 5, we found that the detection mistakes in Table 5 reduce remarkably.

Table 5 Detection mistake of the modified method

e_{max}/mm	$M1$	$M_1 1$	$M_2 1$	$M_3 1$	$M2$	$M_1 2$	$M_2 2$	$M_3 2$	$M5$	$M_1 5$	$M_2 5$	$M_3 5$
10	3	2	1	0	3	2	1	0	3	1	2	0
20	3	2	1	0	3	1	2	0	1	0	1	0
30	3	2	1	0	3	1	1	1	2	0	1	1
40	3	1	2	0	3	1	2	0	2	0	2	0
50	3	0	3	0	2	0	2	0	2	0	2	0
60	3	2	1	0	3	0	3	0	2	0	2	0
70	3	2	1	0	2	1	1	0	3	0	2	1
80	3	0	3	0	3	0	3	0	2	0	2	0
90	3	1	2	0	3	0	3	0	3	0	2	1
100	3	0	3	0	3	0	3	0	2	0	1	1
200	3	0	3	0	3	0	3	0	3	0	2	1
500	3	0	3	0	3	0	3	0	3	0	2	1

注: $M1, M2, M5$: detection mistake of data of 1, 2, 5 stations containing outliers using equation(4), respectively.

The performance features in Table 6 are similar with those of the detection efficiency. Comparing the results of Table 3 with Table 6, the performances of the modified method are better and the number of negative values is less. That is because the detection efficiency is higher and mistakes are less.

Table 6 Performances of the modified method

e_{max}/mm	$R_1 1$	$R_2 1$	$R_3 1$	$R_t 1$	$R_1 2$	$R_2 2$	$R_3 2$	$R_t 2$	$R_1 5$	$R_2 5$	$R_3 5$	$R_t 5$
10	- 0.152	0.124	0	- 0.010	- 0.075	0.076	0	0.007	- 0.009	0.097	0	0.089
20	0.458	0.131	0	0.529	0.493	0.277	0	0.633	0.323	0.307	0.087	0.572
30	0.709	0.007	0	0.711	0.651	0.246	- 0.031	0.728	0.422	0.279	0.077	0.615
40	0.864	- 0.090	0	0.852	0.760	0.166	0	0.799	0.553	0.486	0.107	0.795
50	0.830	- 0.136	0	0.807	0.782	0.110	0	0.806	0.599	0.417	0.112	0.792
60	0.910	- 0.065	0	0.904	0.853	0.166	0	0.878	0.677	0.450	0.065	0.834
70	0.904	- 0.063	0	0.898	0.842	0.283	0	0.886	0.688	0.564	0.072	0.874
80	0.903	- 0.131	0	0.890	0.872	0.166	0	0.893	0.703	0.681	0.159	0.920
90	0.921	- 0.080	0	0.915	0.908	0.216	0.039	0.931	0.710	0.667	0.099	0.913
100	0.946	- 0.280	0	0.931	0.925	0.147	0	0.936	0.737	0.654	0.138	0.921
200	0.974	- 0.328	0	0.965	0.956	0.445	0	0.976	0.898	0.729	0.119	0.976
500	0.995	- 0.696	0	0.992	0.986	0.384	0	0.992	0.982	0.554	- 0.028	0.992
mean	0.772	- 0.134	0	0.782	0.746	0.223	0.001	0.789	0.607	0.490	0.084	0.774

5 Conclusion

The detection capability of robust statistical algorithm incorporating distribution features of the precipitation data is utilized in this paper to detect outlying data in telemetric system. The Tukey fences which are less sensitive to outlying observations are used. The case studied on synthetical data has shown that the proposed method produces reliable detection results. Moreover, the modified method provides better results. That illustrates the importance of distribution features of the precipitation to performance of the method.

It is an integrated problem that resists the extreme outlying rainfall observations in telemetry system, in addition to improving telemetric equipment and reliability of software system, stability and security of flood forecasting system need to be considered. A systemic research should be carried out in future.

References

[1] Barnett V, Lewis T. Outliers in statistical data[M]. UK: John Wiley, 1994.

[2] Han J, Kamber M. Data mining: concepts and techniques[M]. Morgan Kaufmann Publishers, 2001.

[3] Grubbs F E. Procedures for detecting outlying observations in samples[J]. Technometrics, 1969, 11(1): 1-10.

[4] Grubbs F E, Beck G. Extension of sample sizes and percentage points for significance tests of outlying observations[J]. Technometrics, 1972, 4(14): 847-853.

[5] Singh D P. Flood frequency modeling and outliers organic geochemistry[M]. New York: ASCE, 1980.

[6] Hu S Y. Problems with outlier test methods in flood frequency analysis[J]. Journal of Hydrology, 1987, 96(1-4): 375-383.

[7] Spencer C S, McCuen R H. Detection of outliers in pearson type III data[J]. Journal of Hydrologic Engineering, 1996, 1(1): 2-10.

[8] Bounessah M, Atkin B P. An application of exploratory data analysis (EDA) as a robust non-parametric technique for geochemical mapping in a semi-arid climate[J]. Applied Geochemistry, 2003, 18: 1185-1195.

[9] Zhou Q, Li S N, Li X P, *et al*. Detection of outliers and establishment of targets in external quality assessment programs[J]. Clinica Chimica Acta, 2006, 372: 94-97.

[10] Zhou Q, Shen Z Y, Li S N, *et al*. Robust and traditional statistical methods in the establishment of immunoglobulin E target values in external quality essessment program[J]. Clinica Chimica Acta, 2008, 387: 66-70.

[11] Daszykowski M, Kaczmarek K, Heyden Y V, *et al*. Robust statistics in data analysis-A review: basic concepts[J]. Chemometrics and

Intelligent Laboratory Systems, 2007, 85: 203-219.
[12] Grubbs F E. Procedures for detecting outlying observations in samples[J]. Technometrics, 1969, 11(1): 1-10.
[13] Narasimhan S, Mah R. Generalized likelihood ratio method for gross error identification[J]. AIChE J, 1987, 33: 1514-1521.
[14] Bao W M, Qu S M, Li Q S, *et al*. Study of estimation methods of rainfall gauge errors in remote system[J]. Journal of Hydraulic Engineering, 2003, 4: 30-33 (in Chinese).
包为民, 瞿思敏, 李清生, 等. 遥测系统降雨观测误差估计方法研究[J]. 水利学报, 2003, 4: 30-33.

遥测降雨异常值的三步抗差统计探测

赵 超[1,2], 洪华生[2], 朱木兰[1]
(1厦门理工学院水资源环境研究所, 厦门 361005;
2近海海洋环境科学国家重点实验室, 厦门大学环科中心, 厦门 361005)

摘 要 由于遥测降雨系统自身的原因, 遥测降雨资料中常有异常值的出现. 充分利用降雨分布特征以及抗差统计理论, 提出一种三步抗差统计方法探测遥测降雨资料中的异常值. 本方法采用 Tukey fence 统计方法抵御异常值的干扰, 用三步的形式以适应降雨资料的分布特征. 对面平均降雨量进行的分组, 进一步提高方法的探测效率. 数据证明, 新方法的探测效果较好, 且符合水文预报要求.

关键词 遥测系统, 异常值, Tukey fence, 分布特征, 三步抗差探测